Variants
The Journal of the European Society for Textual Scholarship

14 | 2019
Varia

Visualizing Collation Results

Elisa Nury

Electronic version
URL: http://journals.openedition.org/variants/950
DOI: 10.4000/variants.950
ISSN: 1879-6095

Publisher
European Society for Textual Scholarship

Printed version
Number of pages: 75-94
ISSN: 1573-3084

Electronic reference
Elisa Nury, « Visualizing Collation Results », *Variants* [Online], 14 | 2019, Online since 10 July 2019, connection on 12 July 2019. URL : http://journals.openedition.org/variants/950 ; DOI : 10.4000/variants.950

Visualizing Collation Results

Elisa Nury

Abstract: How can the results of automated collation facilitate the analysis of witnesses' relationships? This article introduces the tool PyCoviz, designed to process a collation obtained with CollateX, focusing on the scholarly need to detect shared errors and unique errors. The *Declamations* of Calpurnius Flaccus serves as a case study to show how PyCoviz allows to reproduce an editor's conclusions on the manuscript tradition. While analyzing the collation of Calpurnius' text, this article discusses the difficulty of comparing orthographic differences, and how PyCoviz could be improved to deal with incomplete witnesses, or to visualized editorial uncertainty.

In recent years, the use of computers in the collation workflow has increased, and so has the need to display results of collation in meaningful ways.[1] Collation results are recorded in various digital formats, whether performed by hand or with the help of a tool such as Juxta or CollateX.[2] As more and more texts are available in a digital format, however, and as the efficiency of collation tools improves, computer-supported collation is likely to become the solution of choice in order to collate texts (Prebor 2013, 64).

In this article, I would like to address the issue of visualizing collation results, focusing in particular on results produced by CollateX. What is a good visualization? How can it help the editor to assess the witnesses, their relationship, and to prepare a critical text? What information should be included in the visualization? Collation is admittedly more than just a record of variant readings (Macé et al. 2015, 331). It frequently incorporates additional notes and comments, assessing the certainty of a reading for instance, as well as "paratextual" elements such as changes of pages or folia, gaps, lacunae, and so on. This combination of variants, annotations and paratextual material produces a large amount of complex collation data which is difficult to read and interpret. Therefore, the editor needs to visualize and analyse the collation results as a whole, and not only variant by variant. A good visualization should offer a way to check collation against the actual witnesses, whether they are manuscripts or printed editions. In addition,

1 This research is conducted as part of my PhD research at King's College London, funded by the Swiss National Science Foundation (project no. 155121). Figure 4.6 was developed with the help of my colleague Ginestra Ferraro, UX/UI developer at King's Digital Lab.

2 Juxta Commons: http://juxtacommons.org and CollateX: http://collatex.net [Accessed 4 November 2016].

the editor should be able to interact with the collation to analyse readings and variants in order to evaluate the stemmatic weight of a witness. Collation could be filtered, so as to find patterns of agreements or disagreements between those witnesses, which can indicate how they are related to each other. Visualization and manipulation of collation results are thus essential in order to use collation for further research, such as studying the manuscript tradition and creating a *stemma codicum*.

1. Automated Collation

The role of computers in supporting the collation process goes back to the 1960s, when Dom Jacques Froger (1968) discussed one of the first programs performing collation. Froger was soon followed by other scholars (Gilbert 1973; see also Hockey 1980), but often their software was intended for a limited audience. Collate, an automated collation tool published by Peter Robinson (1994), and the Tübingen System of Text Processing (TUSTEP) developed by Wilhelm Ott (1991), were among the first programs to be used more widely across a range of projects. In 2009, the Gothenburg model was designed in order to improve the complex collation process by dividing it into a series of simpler tasks. Following this model, new collation programs were created, such as CollateX (the successor of Collate) and Juxta. One of the benefits of automated collation is to obtain a collation result in a digital format that can be readily processed for further research.

The creation of a *stemma* with digital methods is one possible application of the results obtained through automated collation, since the collation data is already in a convenient format for processing with a computer. The parallel between manuscript variation and DNA mutation prompted researchers to apply algorithms from evolutionary biology to textual traditions. The phylogenetic approach was adopted with success by Peter Robinson to the Wife of Bath's prologue in Chaucer's *Canterbury Tales* (Barbrook et al. 1998). More recently, the STAM research project carried out by the University of Helsinki has adapted phylogenetic algorithms especially for the purpose of studying textual traditions (Roos and Zou 2011).[3] Their algorithms are available through the Stemmaweb interface of Andrews (2012). The Canterbury Tales project has spurred criticism (Cartlidge 2001) and the debate on the application of phylogenetic methods to textual traditions is still ongoing (Howe et al. 2012; see also Heikkilä 2014, Roelli 2014, and Andrews 2016). The importance for scholars to be able to interact with the collation results was underlined by Andrews and van Zundert (2013). They argue that a static visualization is a barrier to research and that interactivity encourages scholars to be more critical of the results produced by the algorithms of automated collation programs such as CollateX. An interactive visualization of the collation results is therefore an important issue related to automated

[3] See http://cosco.hiit.fi/Projects/STAM/[Accessed 16 February 162017].

collation. The variant graph data model behind collation tools (see Schmidt and Colomb 2009; see also Dekker et al. 2015) serves as visualization in Stemmaweb, Jänicke et al. (2015) proposed a new tool, Traviz, meant to improve the variant graph visalization. The features of Traviz include for instance the use of colours to distinguish witnesses, font-sizes that reveal the frequency of a reading, and a division of the text in lines for better readability. The most advanced tool to correct a collation alignment is the Collation Editor, prepared for the collation of the complex Greek New Testament tradition.[4] The Collation Editor allows numerous interactions, including regularization of orthographic variation, and correction of the alignment.

Beside the variant graph, collation tables are another visualization format. This is the visualization adopted for instance by the Beckett Digital Manuscripts Project or the Digital Mishnah project.[5] While CollateX focuses rather on improving the witnesses' alignment, Juxta places a much stronger emphasis on visualization (Dekker et al. 2015). Juxta offers several visualizations, the Heat Map and a Side-by-Side view, where colours are used to show the places of variation in the text. A histogram is available to show to which degree the witnesses vary across the text. In addition, Juxta also integrates the Versioning Machine among the visualization options, another tool that displays parallel versions of texts encoded in XML TEI with the P5 guidelines and highlights corresponding segments in the different versions.[6] Finally, the TEI Critical Apparatus Toolbox is also designed to display parallel versions when variants are encoded manually by a scholar.[7] Compared to the Versioning Machine, the TEI toolbox provides different features. In particular, the use of colours shows when witnesses agree or not with the readings of a critical text. This can be an important feature for the collation's analysis: the variants marked in orange within the toolbox are considered errors according to a critical text, and errors are essential in the neo-Lachmmann method for text editing.

1.1. Lachmann's Common Errors

To detect relationships between witnesses, many scholars follow the (neo-)Lachmannian method of text editing (Trovato 2014). Here neo-Lachmannism refers to the improvements to Lachmann's method brought by Giorgio Pasquali and other Italian scholars, who took Bédier's criticism into account and incorporated the study of the textual tradition and material documents (the manuscripts themselves) to the creation of stemmata (Pasquali

---

4 The Collation Editor is a tool produced by The Institute for Textual Scholarship and Electronic Editing (ITSEE) at the University of Birmingham, as part of the Workspace for Collaborative Editing: http://vmrcre.org/ [Accessed 4 November 2016].

5 Samuel Beckett. Digital Manuscript Project: http://www.beckettarchive.org/ and Digital Mishnah: Developing a Digital Edition of the Mishnah: http://www.digitalmishnah.org/ [Accessed 4 November 2016].

6 Versioning Machine: http://v-machine.org/ [Accessed 2 December 2016].

7 The TEI critical apparatus Toolbox: [Accessed 28 June 2016].

1952). Lachmann's method focuses on common errors shared by a group of witnesses in order to postulate relationships between those witnesses: witnesses are likely to be related if they (1) agree on readings that (2) they do not share with the other witnesses, and especially (3) those that agree in errors (i.e. they share readings that have no manuscript authority). A reading with manuscript authority is "a reading that may have reached us through a continuous sequence of accurate copies of what the author wrote back in antiquity and may therefore be authentic and (by definition) right" (Damon 2016, 202–3). In sum, witnesses may be related when they share readings that do not represent the original text and that are absent from other witnesses. Other readings of interest in Lachmann's method are "unique errors", which are errors found only in one witness. Between two related witnesses that share common errors, if one has in addition unique errors that could not have been easily corrected, it can be concluded that this witness is a direct descendant of the other (West 1973, 33). Being able to find common errors or unique errors in the collation results would therefore be especially useful for a scholar preparing a critical edition.

The visualizations described above are mostly linear: the reader must follow the text word by word and it is difficult to select only variants of interest, such as common readings or unique readings. A new perspective may be required in order to fulfill the need of editors: the ability to filter the collation and select agreements between witnesses or unique readings. In addition, the collation tables will usually show only plain text from the witnesses and omit the paratextual elements that could be useful to an editor. Enhancements have been proposed to improve the basic table output of CollateX, for instance with colours to indicate the places where a variation occurs: in the Digital Mishnah demo, variant locations are highlighted in grey. Another example is the Beckett Digital Manuscripts Project, where deletions are represented with strikethrough and additions with superscript letters.[8] However, other elements are still missing from those helpful visualizations, such as, for instance, the changes of folia mentioned earlier, or other types of paratextual and editorial annotations. The reason for recording folia changes is mainly for checking purposes. If the editor or a reader wants to check the accuracy of the transcription for a particular reading, it will be much easier to find the reading back in the manuscript knowing the folio where it appears.

In the context of the *Declamations* of Calpurnius Flaccus, the need for visualizing groups of witnesses with shared readings, or unique readings, has become necessary to analyse the manuscript tradition and the relationship of the *editio princeps* with other manuscripts. Therefore, I have developed a tool to respond to this need: a user interface which makes it possible to filter the collation results in order to find groups of witnesses that agree with one another and not with others or to find the readings unique to one witness. The results of these searches

[8] See the news update of 17 September 2014 at http://www.beckettarchive.org/news.jsp [Accessed 4 November 2016].

are then displayed within a collation table that incorporates some paratextual elements. This tool will be described in more detail in Section 3.

2. Case study: Calpurnius Flaccus

The case study to which the visualization tool was applied is the *Declamations* of Calpurnius Flaccus, a classical literary text in Latin from the second century AD. Declamation was originally a Greek practice which was adopted in the Roman world. The production of declamations started as school exercises meant for students to practice their rhetorical skills and the art of public speaking. The most difficult of those exercises were the *controversiae*, legal speeches in fictitious court cases. Given a situation of conflict (the theme) and a set of laws, the students had to play the part of a lawyer and learn to defend both parties. The characters portrayed in declamations are everyday members of society: fathers and sons, mothers-in-law, young women and rapists, rich and poor enemies, deserters and war heroes. From the personality associated with those anonymous persons, the purpose of the declamation exercise is to build a convincing plea with the help of witty traits, the *sententiae*. Declamations evolved also in a literary genre of its own, with performances from well-known rhetors for the public entertainment (Sussman 1994). The corpus of *Declamations* from Calpurnius Flaccus is a collection of fifty-three declamation extracts: besides the titles and themes, we do not possess complete speeches but only the most noteworthy *sententiae* of the author. Little is known about Calpurnius, except for his name; Sussman (1994, 6) places his work in the second century based on his style.

2.1. The Manuscripts

The text is transmitted by five manuscripts. The older surviving manuscript is codex Montepessulanus H 126 (A), held in the Bibliothèque Universitaire de Médecine in Montpellier. Manuscript A is a very valuable witness, but unfortunately badly damaged. Only the last folio provides us with the first six declamations, which are very difficult to read due to dark stains on the page. A large part of the folio has become completely illegible. Because of its fragmentary and lacunose character, the manuscript was omitted in this visualization example. The treatment of this manuscript raises issues that will be discussed below (Section 5).

A lost manuscript (X) appears in the correspondence of Humanist scholars: this manuscript is likely the source of two manuscripts from the fifteenth century, codex Monacensis Latinus 309 (B) and codex Chigianus Latinus H VIII 261 (C), held respectively in Munich's Bayerische Staatsbibliothek and in the Vatican Library. The tradition is completed by two other manuscripts from the sixteenth century, codex Monacensis Latinus 316 (M) in Munich and Bernensis Latinus 149 (N) in Bern. Also in the sixteenth century, the French scholar Pierre Pithou published the *editio princeps* in 1580, and reprinted it fourteen years later in 1594. The reprinted edition was used here in the collation because it is digitized and more easily available.[9] Nevertheless, the two versions of 1580 and 1594 have been compared and three important differences, other than abbreviations, were noted: in Declamation 1, *filii mei mortem* was replaced with *mortem mei filii* (Pithou 1594, 383); in Declamation 21, *possem* was replaced with *posse* (Pithou 1594, 400); finally, in Declamation 34, *medius* was replaced with *melius* (Pithou 1594, 409).

The critical edition of Calpurnius Flaccus published by Lennart Håkanson in 1978 is also included in the collation. Håkanson's text is still the best critical edition available, complete with a comprehensive critical apparatus. In his introduction, Håkanson gives a detailed analysis of the relationships between the different manuscripts and Pithou's edition. The next section will summarize Håkanson's conclusions, and Section 4 will then examine how the visualization tool could be used to analyse the tradition of Calpurnius Flaccus.

2.2. The Stemma

In the preface to the *Declamations*, Håkanson describes in detail the reasoning process behind the stemma construction. He gives a practical example of the application of Lachmann's method to a Latin literary text. Here is a summary of how Håkanson established his stemma. We will compare his results with the ones we obtain through our automatic collation and interactive visualization later.

Håkanson shows first that all five manuscripts descend from a common archetype, since they share a few errors (Håkanson 1978 VI). He postulates then that A and X, the lost hyparchetype of BCMN, form two distinct branches of the stemma. However, the text of Calpurnius is too short in manuscript A to prove this point. Instead, Håkanson relies on other texts transmitted in the manuscripts A, B, and C.

Next, Håkanson proceeds to analyse relationships between BCMN: BMN have some errors in common which are absent from C. Therefore, the stemma is divided again in two branches stemming from X, with C on one side and BMN on the other side (Håkanson 1978 VII-VIII). Manuscripts M and N are separated from B by errors that they have in common (Håkanson 1978 VIII), and interpolations which have been introduced in their exemplar by an unknown witness Y (Håkanson 1978 IX). Furthermore, there was an exchange of readings between manuscripts B and N. A few readings from N have been added in the margins of B by a second hand (B2): *vel* (22.4), *remittitur* (24.11) and *in vita liberis* (25.17) (Håkanson 1978 XII).

---

[9] The edition is available on the portal e-rara, a collection of digitized printed book from Swiss universities: http://www.e-rara.ch/gep_g/content/titleinfo/976587 [Accessed 22 February 2017].

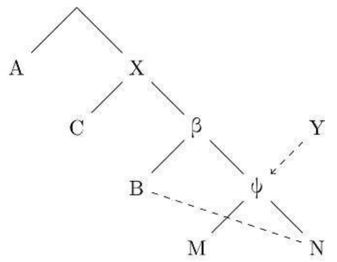


Figure 4.1: Calpurnius Stemma (Håkanson 1978, V).

The complete stemma of Calpurnius Flaccus is showed in Figure 4.1. After sorting out the manuscripts' relationships, Håkanson examines how the *editio princeps* is related to the manuscripts. The *editio princeps* of Pithou is based in part on manuscript A. Since it is damaged and incomplete, Pithou had to rely mostly on another manuscript which he referred to as the "Italian exemplar". Pithou does not give much detail about this codex, but the only manuscript that we know for certain to have been in Italy is manuscript C, now held in the Vatican Library. However, both M and N are written in italics, an indication of a potential Italian origin. In fact, Håkanson argues for N to be the Italian manuscript mentioned by Pithou, because readings unique to N were adopted by Pithou (Håkanson 1978, XIII).

Furthermore, Jacques Bongars, the last owner of manuscript N before its acquisition by the Bern Burgerbibliothek, was in close contact with Pierre Pithou (Banderier 2009, 397). Bongars could have shared the manuscript with Pithou.

Was N really the Italian manuscript that Pithou used for his edition, or is it possible that N was copied from Pithou's edition? Both the edition and the manuscript are from the late sixteenth century and it is not possible to ascertain which one is oldest. In his apparatus, Pithou quotes the reading "*miseriae nostre aur*". which is not present in any known manuscript of Calpurnius Flaccus.[10] The other witnesses read *gemitum miseri aures tuae* (BCMN) or *miseri mei gemitum aures tuae* (A) at this point in the text. Is it only a mistake from Pithou or could this be an indication that N was not Pithou's Italian manuscript? How could the relationship of Pithou's edition with the other manuscripts be examined with the help of the tool and collation tables? In Section 4.2, we will see how this visualization tool is used and how it can help us to check Håkanson's conclusions on the tradition of Calpurnius Flaccus.

[10] See Pithou's apparatus: http://www.e-rara.ch/gep_g/content/pageview/1099004 [Accessed 23 February 2017].

3. Visualization tool: PyCoviz

In the Calpurnius case study, four manuscripts have been considered for visualization: manuscripts B, C, M and N. Each manuscript is split into two witnesses according to the different hands which wrote the text. For example, B1 is the code for the first hand of manuscript B and B2 is the code for the second hand responsible for corrections to the text of B1. There are also two editions in the collation, Pithou's edition of 1594 (P1594) and Håkanson's edition of 1978 (LH). There are thus ten witnesses in total, which have been transcribed following the XML TEI P5 guidelines, and then prepared to be collated with CollateX, version 1.7.1.

The method of visualization for CollateX's results consists of two aspects: first, a Jupyter notebook, PyCoviz (for Python Collation Visualization), where the editor can interact with the collation results through a Python script (for instance to select agreements between a group of witnesses against another group) to make small corrections in the alignment or to search for readings.[11] Second, a collation table in HTML format, with additional information such as reading's location in the manuscripts or links to digital facsimile.

A Jupyter notebook is a document format that combines computer code with prose descriptions and that is accessible through a web browser. Notebooks are especially designed to share or publish executable code, which makes it a well-suited format for sharing the Python script developed for collation visualization (Kluyver et al. 2016; see also VanderPlas 2016). Since CollateX is distributed in Python, this was the preferred coding language for the customization of a Jupyter notebook. The combination of code and prose explanations should help to make the notebook accessible to scholars with little knowledge of coding or Python.

The addition of widgets into the Jupyter notebook offers an interactive way to explore collation results. Widgets are components of a user interface such as buttons, text boxes, and so on.[12] The results from CollateX are uploaded in PyCoviz and then transformed into Python data format. Then the collation data can be manipulated through various interactions: first, a few functions allow for modifying or correcting the collation if the current alignment is not satisfactory. Second, the collation can be filtered in order to find agreements between selected witnesses. Finally, it is possible to search the collation results and to clarify a reading by displaying all its properties. According to Andrews and van Zundert (2013) essential interaction requirements should be twofold:

1. Alter or correct the collation alignment (this includes combining or splitting words into readings when necessary).
2. Annotate variants about how they are related to each other.

---

[11] https://jupyter.org/ [Accessed 8 November 2016].

[12] For more information about Jupyter's widget, see the documentation: https://ipywidgets.readthedocs.io/en/latest/examples/Widget%20Basics.html [Accessed 14 December 2016].

The first aspect is fully implemented in the notebook, but readings can be annotated only individually. However, it is arguable that the annotation of readings may be better represented in an external database such as described by Spadini (2015). In addition to those two kind of interactions, the possibility to search for agreements of witnesses offers another option to analyse collation results.

3.1. Collation Format

By default, CollateX can take as input plain text transcriptions of the witnesses to collate. The texts will be split into "tokens", smaller units of text, at white spaces. This is the tokenization stage.[13] The collation is then performed on these tokens, which are usually the words of the text. However, it is also possible to "pre-tokenize" the transcriptions and divide texts into tokens according to the user's needs. The tokens may then be provided as input to CollateX, using for instance the JSON format, instead of plain text. The JSON format allows one to record not only the plain text words (t), but also a normalized form of the word (n) or other properties. There is no limitation to the token properties that can be added: they will simply be ignored during the collation stage, but still be available in the end results. In order to integrate folio location, links to digital images and editorial comments, the TEI transcriptions of Calpurnius Flaccus were transformed into pre-tokenized JSON. The tokens include, beside the (t) and (n) properties, a (location) property, eventually a (link) to a digital facsimile and/or a (note) property.

During collation, the properties of location, link and note are ignored, and only tokens (t) and their normalized (n) forms will be compared. Normalized forms are compared first. In the absence of (n), CollateX will then compare tokens in their original form, namely tokens (t) which represent the words as they appear in the witnesses. For the pre-tokenization of Calpurnius, abbreviations were expanded and punctuation was not included. This is not to say that punctuation marks or abbreviations are not important; in fact, a change in punctuation can considerably affect the meaning of the text and abbreviations are often a source of errors (West 1973, 27). However, in a Classical Latin tradition such as the one of Calpurnius Flaccus, both abbreviations and punctuation marks are more characteristic of the scribes who copied the manuscripts than of the author's language, and for this reason have not been collated for the text discussed here. Tokens may be normalized in different ways, but in our case it derives directly form the TEI transcription of each witness in which spelling variations are normalized via the use of TEI elements `<orig>` and `<reg>` (TEI Guidelines, chapter 3.4.2). The `<orig>` element provides the content of token (t), while `<reg>` provides the normalized form (n) (for instance the words *foemina* or *fęmina* would be normalized to *femina*). In addition, during the transformation from the TEI to the JSON format, uppercase letters are all normalized into low-

[13] See CollateX documentation: http://collatex.net/doc/#input [Accessed 4 November 2016].

ercase, accented letters in Pithou's edition are replaced by their non-accented counterparts, and ampersand symbols (&) are replaced by "et". This normalization process can lead to useful collation results from CollateX (Dekker et al. 2015, 4). Moreover, for visualization purposes, it may also be desirable to ignore trivial variations that cannot be considered as errors in the context of neo-Lachmann's method. CollateX offers various output formats of the collation results. For this visualization, the JSON output format was preferred: it is a format which can be easily manipulated with Python in order to produce our visualization. The only missing pieces of information in the JSON output are the indication of transpositions, i.e., segments in which the order of words do not coincide between the witnesses.[14] However, transpositions are not a major issue within the text of Calpurnius Flaccus, and therefore the JSON output has served well in this case study. The JSON output comes in two different forms: with consecutive matching tokens joined into segments, i.e. consecutive words which are considered equivalent are aligned together in a single row of the collation table — see Figure 4.2(a). The second option is to separate each token into a different row. Figure 4.2(b) shows how the second option can be problematic because differences in word division may lead to a confusing output (in the example, *ad te* should be aligned with *ante*, and *rem publicam* with *rempublicam*). The JSON output with consecutive matching tokens joined into segments was used for the text of Calpurnius, and next section will discuss the two outputs and their issues.

B1	B2	C1	C2	LH	M1	M2	N1	N2	P1594	ID
gloria non	gloria non	gloria non	gloria non	gloria Non	gloria non	gloria non	gloria non	gloria non	gloria Non	
ad te	apte	ad te	ad te	ante	ante	ante	ante	ante	ante	492
rem publicam	rem publicam	rem publicam	rem publicam	rem publicam	rem publicam	rem publicam	rempublicam	rempublicam	Rempublicam	493

(a) Consecutive matching tokens joined into segments.

non	non	non	non	Non	non	non	non	non	Non	2584
ad	apte	ad	ad	ante	ante	ante	ante	ante	ante	2585
te		te	te				rempublicam	rempublicam	Rempublicam	2586
rem	rem	rem	rem	rem	rem	rem				2587
publicam	publicam	publicam	publicam	publicam	publicam	publicam				2588

(b) Tokens as separate rows in the collation table.

Figure 4.2: CollateX's JSON output options

3.2. The issue of normalized tokens

The collation tables in PyCoviz are all made by comparing normalized forms of tokens whenever possible. In practice, the orthographical differences are thus excluded from the collation tables. However useful it may seem, the use of normalized forms for the analysis of collation results may be problematic in certain

[14] Additional support to visualize transpositions in a table format will be added in the future, according to Collatex's documentation: https://collatex.net/doc/#json-output [Accessed 5 March 2017].

circumstances. Comparing only normalized forms may hide variant readings that could be considered significant to an editor. In folio 83^{r}, manuscript C reads *liniamentis*, while the other witnesses read *lineamentis*. This is a purely orthographic difference, and as such, does not appear in the collation table as a place of variation since the comparison is done on the normalized tokens. A user of the Jupyter notebook, selecting witnesses in order to find their (dis)agreements, would not see this row in the results for any combination of groups of witnesses. However, Håkanson (1978, 7) included this orthographic difference in his critical apparatus: therefore this reading was considered to be somehow significant to Håkanson, but according to the method applied here, the reading would not be visible while using the notebook to find agreements between witnesses. For this reason, the notebook also provides a set of functions to compare readings in their original forms, i.e. using tokens (t) instead of their normalized forms (n). Nevertheless, this approach is not a satisfactory solution, because CollateX's results used here have consecutive matching tokens joined into segments. As a consequence, some large chunks of texts are combined into a single cell of the table when there is no difference between normalized forms. Even if there is an orthographic difference, it could be hard to spot it in the middle of a long block of text: the word *liniamentis* mentioned above appears in the middle of a 37-word reading that shows other orthographic variations (see Figure 4.3). Comparing orthographic differences is therefore difficult.

B1	B2	C1	C2	LH	M1	M2	N1	N2	P1594	ID
nunquam iudices contra istum tutior veni quicquid auferri potuit amisi soli omnium torti sunt donec mentirentur ita laniatos miseros ita confusis lineamentis proiici iussit ut iam nec pater posset agnoscere miramini si ab hostibus sepulti sunt nullos	nunquam iudices contra istum tutior veni quicquid auferri potuit amisi soli omnium torti sunt donec mentirentur ita laniatos miseros ita confusis lineamentis proiici iussit ut iam nec pater posset agnoscere miramini si ab hostibus sepulti sunt nullos	nunquam iudices contra istum tutior veni quicquid auferri potuit amisi soli omnium torti sunt donec mentirentur ita laniatos miseros ita confusis liniamentis proici iussit ut iam nec pater posset agnoscere miramini si ab hostibus sepulti sunt nullos	nunquam iudices contra istum tutior veni quicquid auferri potuit amisi soli omnium torti sunt donec mentirentur ita laniatos miseros ita confusis liniamentis proici iussit ut iam nec pater posset agnoscere miramini si ab hostibus sepulti sunt nullos	Numquam iudices contra istum tutior veni quicquid auferri potuit amisi Soli omnium torti sunt donec mentirentur Ita laniatos miseros ita confusis lineamentis proici iussit ut iam nec pater posset agnoscere miramini si ab hostibus sepulti sunt nullos	numquam iudices contra istum tutior veni quicquid auferri potuit amisi soli omnium torti sunt donec mentirentur ita laniatos miseros ita confusis lineamentis proici iussit ut iam nec pater posset agnoscere miramini si ab hostibus sepulti sunt nullos	numquam iudices contra istum tutior veni quicquid auferri potuit amisi soli omnium torti sunt donec mentirentur ita laniatos miseros ita confusis lineamentis proici iussit ut iam nec pater posset agnoscere miramini si ab hostibus sepulti sunt nullos	nunquam iudices contra istum tutior veni quicquid auferri potuit amisi soli omnium torti sunt donec mentirentur ita laniatos miseros ita confusis lineamentis proiici iussit ut iam nec pater posset agnoscere miramini si ab hostibus sepulti sunt nullos	nunquam iudices contra istum tutior veni quicquid auferri potuit amisi soli omnium torti sunt donec mentirentur ita laniatos miseros ita confusis lineamentis proiici iussit ut iam nec pater posset agnoscere miramini si ab hostibus sepulti sunt nullos	Nunquam iudices contra istum tutior veni quicquid auferri potuit amisi soli omnium torti sunt donec mentirentur Ita laniatos miseros ita confusis lineamentis proiici iussit ut iam nec pater posset agnoscere Miramini si ab hostibus sepulti sunt nullos	201

Figure 4.3: Liniamentis in the collation results.

CollateX does also provide results where consecutive matching tokens are not joined into segments, but in this case the information that some groups of tokens should be considered together is lost, especially when one token of one witness matches with several tokens in another witness. It is a fairly common situation in Latin, since texts used to be written in the *scriptio continua* style, without word division. Inconsistencies in word division are quite frequent in

Calpurnius: *rempublicam* versus *rem publicam*, *contradicit* versus *contra dicit*, and so on. Some cases are more complex than just difference in word division, such as *verberantibus* in B1 corrected into *verbera cibus* by B2 (f. 148^{v}), where one word matches two different words of another witness. In fact, these situations were so frequent that they were not normalized in the TEI transcriptions. Rather, it was decided to compare normalized forms without spaces in between words, so that word division would not be considered a variant: PyCoviz will consider that witnesses with the readings *eius demet* (BC) and *eiusdem et* (LH, P1594) agree together, because white spaces are not included. Hence the comparison will be made on the reading *eiusdemet* which is equivalent in the four witnesses. There are so many instances of one-to-many matching tokens that correcting CollateX results, when matching tokens are not joined into segments, would not be worth the effort. The example of Figure 4.2(b) showed how the alignment of tokens is much less accurate when each token is in a separate row. Any attempt to toggle between the two results of CollateX, with and without joint segments, is bound to be difficult because of these many places where there is no one-to-one matching token. One possibility would be to normalize tokens only after the collation is done. While it may be achievable for the *Declamations* or other texts with a limited amount of orthographic variation, it may not be desirable for other traditions, such as medieval traditions where there are countless orthographic variants. The best solution may still be to highlight the orthographic differences within the collation table, while using the CollateX results with consecutive matching tokens joined into segments.

4. Exploring the Collation of Calpurnius Flaccus

4.1. Corrections

Before analysing the collation and searching for witnesses' agreements, it may be necessary to inspect the results and eventually correct them. The corrections described in this section were made with the help of widgets within PyCoviz and demonstrate how the Jupyter notebook may be useful to adapt the collation results according to the user's need. In some cases, CollateX's result was clearly incorrect: for instance, *proximi*, a conjecture by Håkanson (1978, 1.2), was not aligned with the reading *proximae* present in the other witnesses. The token was thus moved in order to be in the right place. The alignment could also be refined, even if it is not actually wrong: the reading *luxuriosum ob amorem* in Håkanson (1978, 27.13) is aligned with *ob amore* in the other witnesses. The user might want to split the reading in two, so that the conjecture *luxuriosum* printed by Håkanson does not match with other readings, whereas *ob amorem* would be aligned with *ab amore* present in the other witnesses.

Other situations are less obvious, and the final alignment might be debatable. For instance, the reading *invitabo nisi* in witnesses B1 (f. 155^{v}), C1 and C2 (f. 87^{r}) is closely related to *in vita bonis* found in M1 and M2 (f. 13^{v}). The variant appears to originate mainly from a word division issue: it seems thus that the readings might belong to the same row. However, witnesses N1, N2, B2 and LH read *in vita liberis*. Therefore, *bonis* was ultimately aligned with *liberis* and *nisi*, even if that means that the letters *-bo* in *invitabo* are not aligned anymore with *bo-* in *bonis* (see Figure 4.4). This alignment makes it possible to show the agreement of the witnesses with the reading *in vita*, however another user might decide that the whole group of words should be aligned together: *invitabo nisi* aligned with *in vita bonis* and *in vita liberis*.

B1	B2	C1	C2	LH	M1	M2	N1	N2	P1594	ID
invitabo	in vita	invitabo	invitabo	in vita	in vita	in vita	in vita	in vita	in vita	847
nisi	liberis	nisi	nisi	bonis	bonis	bonis	liberis	liberis	liberis	848

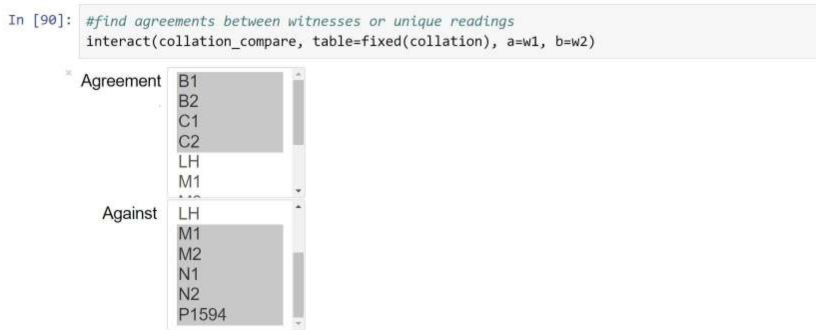
Figure 4.4: Alignment of *invitabo nisi* with *in vita bonis* and *in vita liberis*.

Within the *Declamations* of Calpurnius Flaccus, a total of 171 corrections have been made to the alignment provided by CollateX: these corrections include moving 157 tokens, adding two rows and deleting twelve. Out of the 68,914 tokens from ten witnesses, only 157 were not properly aligned by the algorithm, which represent around 0.23 percent of the total number of tokens. Even considering that some errors or mismatches in the alignment escaped correction, the percentage is unlikely to get higher than 0.5. This low percentage may attest to CollateX's efficiency. However, the ten witnesses of this case study represent a small textual tradition, especially since four of the ten witnesses are actually artificially created by attributing the status of witness to corrections of second hands and thus are very similar to the first hand. In addition, only few transpositions can be found in Calpurnius, and the transpositions present in the text usually involve no more than two or three words.[15] Therefore it is likely that, for texts with a more complex tradition, a higher percentage of errors may arise.

4.2. Analysis

Once the user is satisfied with the collation alignment, it is possible to explore the results. As discussed above (Section 1.1), it could be useful to search for and visualize agreements among witnesses in order to find patterns that indicate a relationship. To this end, PyCoviz provides an interactive widget which allows the user to find the agreements of groups of witnesses — see Figure 4.5(a).

15 Transposition detection is a difficult problem for collation algorithms (Dekker et al. 2015, 5).



(a) Selecting agreements with widgets.

B1	B2	C1	C2	LH	M1	M2	N1	N2	P1594	ID
rhetorum minorum	rhetorum minorum	rhetorum minorum	rhetorum minorum						Rhetorum minorum	3
irascitur	irascitur	irascitur	irascitur	irascitur	irascatur	irascatur	irascatur	irascatur	irascitur	129
odit	odit	odit	odit	odit					odit	134
					odit	odit	odit	odit		136
infamet	infamet	infamet	infamet	infamet	infament	infament	infament	infament	infamet	191
quid	quid	quid	quid	quid	quod	quod	quod	quod	quid	255
queri	queri	quẹri	quẹri	queri	quaeri	quaeri	quaeri	quaeri	queri	350
					cuncta	cuncta	cuncta	cuncta		435
eius demet	eius demet	eius dem&	eius dem&	eiusdem et	eiusdem met	eiusdem met	eiusdem met	eiusdem met	eiusdem &	786

(b) Agreements of BC and P1594, against MN.

Figure 4.5: View of the Jupyter notebook PyCoviz.

The cells of the collation tables displayed in the notebook are coloured in red and green, according to their relationship to the reading of a base text; in our case, the base text is the edition of Håkanson (see the examples above, for instance Figure 4.5(a)). Each witness is in a separate column; the last column (ID) represents a way to identify rows in the table. Since the base text is the text printed in the edition of Håkanson, the readings he rejected as errors are shown in red, while the readings he accepted as genuine are shown in green. The colour pattern would of course be different if another edition had been selected as the base text. Using the red and green colour scheme was inspired in part by Stemmaweb, a tool for creating stemmata with computer algorithms (Andrews 2012).[16]

In the case of Calpurnius Flaccus, PyCoviz lets us reproduce Håkanson's conclusions by filtering the collation to find shared errors among the witnesses.

[16] Stemmaweb makes it possible to visualize collation tables where readings are highlighted in green or red. A green highlight means that the reading is consistent with a given stemma hypothesis. A red highlight means that the reading is not consistent with that same stemma hypothesis. In a similar way, the collation table highlights readings that are consistent or not with a "text hypothesis", the text that was selected as a base text.

For instance, the *editio princeps* P1594 has close to forty errors shared with M and N, but only one error shared with B or C. Moreover this single error appears in the *incipit*: it is displayed as an error only because Håkanson did not include the *incipit* in his text. Further use of the Agreement widget can show that, according to Håkanson, P1594 has more errors in common with N than M. Finally, it is possible to see that P1594 has many more unique errors than N. This is how we can reproduce Håkanson's conclusion that N is most likely the 'Italian manuscript' which served as an exemplar to Pithoeus for his *editio princeps*.

4.3. Collation Tables

B1	B2	C1	C2	LH	M1	M2	N1	N2	P1594	ID
excerpta	excerpta	excerpta	excerpta							1
					contra matrem	contra matrem	contra matrem	contra matrem	contra matrem	16
					contra matronam	contra matronam	contra matronam	contra matronam	contra matronam	24
					pro milite	pro milite	pro milite	pro milite	pro milite	65
ↄo	ↄo	ↄo	ↄo		o	o	o	o		67
Note: Unknown abbreviation. Normalized form supplied by Lehnert. **148r:8**	**148r:8**	82r:18	82r:18		**2r:7**	**2r:7**	244v:28	244v:28		
verginius	verginius	verginius	verginius	Verginius	virginius	virginius	virginius	virginius	Virginius	76
					pro parricida	pro parricida	pro parricida	pro parricida	pro parricida	84

Figure 4.6: Collation table with paratextual elements.

PyCoviz can only display simple HTML tables. The tables show words in their original form, and colours to distinguish between errors and true readings. However, in Section 1.1 I outlined the importance of paratextual elements. As these elements are included in a token's properties, they can be made available in a more complex table format. Figure 4.6 shows an example of an improved table visualization, where notes are included as well as the location of tokens in the witnesses. When possible, the location links to a digital facsimile of the page. Both notes and location can be hidden from the table by clicking on the symbols (i) and (). Thus the readability of the collation is not impaired by the extensive additional data.

5. PyCoviz Issues and Further Development

The Jupyter notebook has advantages: it provides a user interface with interactive widgets and explanations along with the code. It is a great tool to share code and to make it available for other scholars to examine and review, and adapt it to their own needs if they wish to do so. The notebook was designed with reuse in mind. Consequently, it should be able to run with any CollateX result that has at least tokens (t). All other features of the tokens are optional. It is not necessary either to choose a base text and view readings as errors or genuine readings. There is no limitation regarding the number of witnesses present in the collation.

Nonetheless, there are a few inconvenient coding aspects with a Jupyter notebook. Version control is notably challenging because the notebook is saved in JSON format, which is not practical for visualizing changes in the code or in the code's output. The presentation of code in blocs, without line numbers, can make it difficult to refer to a precise portion of code. Finally, it may be complicated to keep track of which version of the code produced which collation tables, based on which collation file (obtained from which version of CollateX). It is not an issue related to Jupyter notebooks alone, but to any output obtained with computational method. This kind of information is crucial for the reproducibility of materials obtained with computational methods, as well as for quoting those results in publications. The latest release of PyCoviz was prepared with Jupyter Notebook v4.3 and ipywidget v5.0. The collation of Calpurnius was obtained with CollateX v1.5.

There are still many technical issues that could be subject to improvement in PyCoviz. One of those issues is dealing with incomplete witnesses such as manuscript A in Calpurnius Flaccus. The collation results did not distinguish between text that is missing due to illegible text, or because folia are missing. In other cases, there are empty cells in the collation table when the text is absent for one witness but not in others: for instance, declamation 45 is present only in manuscript C and in Håkanson's edition, but not in other witnesses, although there is no damage to the manuscripts, and the text is perfectly legible. The missing folia of A were represented by empty cells in the collation table. However, this would skew the results when looking for agreements. For instance, looking for agreement between A and other witnesses would return the rows where A is non-existent, and other witnesses happen to have an empty cell. The same problem appears when comparing other witnesses against A: the search would yield many results of agreements against A, when the witnesses agree together and A has an empty cell. This does not mean that the text of A was really different. We don't know because the manuscript evidence simply does not exist for A at this point of the text. The collation table should therefore indicate explicitly the state of A, whether the manuscript is extant or not. When searching for agreements between A and other witnesses, it should be possible to ignore those rows where the folia are missing in A.

Another possible improvement would be to allow to choose a lemma among the available readings in a row and that way create a new base text instead of choosing one of the existing witnesses. It should also be possible to visualize uncertainty: it is not always possible to decide among several variants which is the true reading and which are the errors. Although editors have to choose one to print in their critical text, they often express their uncertainty in the critical apparatus. For instance, the use of a third colour such as yellow could express the editor's uncertainty.

6. Conclusion

The two examples of visualization described above, PyCoviz and the collation table, demonstrate how to make use of collation in an electronic format for further research. The case study of Calpurnius Flaccus was based on a particular collation format obtained with CollateX. Nevertheless, the methodology behind the creation of this visualization should be applicable as well to collation prepared manually and in different digital formats. Collation results obtained with CollateX can benefit from the use of pre-tokenized JSON, as it was already done by other projects such as the Beckett Digital Manuscript Project. However, it is possible to integrate more information into the collation with JSON tokens: elements such as location of a word in the manuscript, or editorial comments, are important aspects of collating texts and there is no reason to discard them in a computer-supported collation. As shown in the collation table, the use of a few symbols allows for making those elements easily available without overcrowding the results. The use of colours is a straightforward way to reveal groups of witnesses which agree with one another and thus help draw conclusions about the manuscript tradition. However, it is also important to keep in mind how the comparison is done and how normalization influences the collation tables obtained when searching for agreements of witnesses. Errors and genuine readings should also be carefully considered as the product of an editor's judgement on the text, in this case the decisions of Håkanson. The use of a different base text allows for different interpretation of the text to be visualized. The collation table and Jupyter notebook presented here will hopefully provide suggestions on how to make available the extra material that is not yet fully exploited in collation visualizations.

Bibliography

Andrews, Tara. 2012. "Stemmaweb — A Collection of Tools for Analysis of Collated Texts". <https://stemmaweb.net> [Accessed 5 November 2016].

——. 2016. "Analysis of Variation Significance in Artificial Traditions Using Stemmaweb". *Digital Scholarship in the Humanities*, 31(3), pp. 523-39.

Andrews, Tara and Joris van Zundert. 2013. "An Interactive Interface for Text Variant Graph Models". In *Digital Humanities 2013: Conference Abstracts, University of Nebraska-Lincoln USA, 16-19 July 2013*, pp. 89–91.

Banderier, Gilles. 2009. "Bâle et la famille Pithou: Contribution à l'étude des rapports intellectuels entre Bâle et la France au XVIe siècle Bâle et la famille Pithou". *Revue Suisse d'histoire*, 59, pp. 387-409.

Barbrook, Adrian C., et al. 1998. "The phylogeny of the Canterbury Tales". *Nature*, 394 (27 August), p. 839.

Calpurnius Flaccus. 1978. *Calpurnii Flacci Declamationum Excerpta*. Ed. Håkanson, Lennart. Calpurnii Flacci Declamationum Excerpta. Stutgardiae: Teubner.

——. 1994. *The Declamations of Calpurnius Flaccus: Text, Translation, and Commentary*. Ed. Sussman, Lewis A. Leiden: New York: E. J. Brill.

Cartlidge, Neil. 2001. "The Canterbury Tales and Cladistics". *Neuphilologische Mitteilungen*, 102, pp. 135–150.

Damon, Cynthia. 2016. "Beyond Variants: Some Digital Desiderata for the Critical Apparatus of Ancient Greek and Latin Texts". In Matthew James Driscoll and Elena Pierazzo (eds.), *Digital Scholarly Editing: Theories and Practices*. Cambridge: Open Book Publishers, pp. 201–218

Dekker, Ronald Haentjens et al. 2015. "Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project". *Literary and Linguistic Computing*, 30(3), 452-70.

Froger, Jaques. 1968. *La critique des textes et son automatisation*. Paris: Dunod.

Gilbert, Penny. 1973. "Automatic Collation: A Technique for Medieval Texts", *Computers and the Humanities*, 7(3), pp. 139–147.

Heikkilä, Tuomas. 2014. "The Possibilities and challenges of computer-assisted stemmatology: the example of Vita et miracula s. Symeonis Treverensis". In Tara Andrews and Caroline Macé (eds.), *The Analysis of Ancient and Medieval Texts and Manuscripts: Digital Methods*. Turnhout: Brepols, pp. 19–42.

Hockey, Susan. 1980. *A Guide to Computer Applications in the Humanities*. London: Duckworth.

Howe, Christopher J., Ruth Connolly, and Heather F. Windram. 2012. "Responding to Criticisms of Phylogenetic Methods in Stemmatology". *SEL Studies in English Literature 1500-1900*, 52(1), pp. 51–67.

Jänicke, Stefan, Marco Büchler, and Gerik Scheuermann. 2014. "Improving the Layout for Text Variant Graphs". In *VisLR: Visualization as Added Value in the Development, Use and Evaluation of Language Resources*, pp. 41–48.

Jänicke, Stefan, et al. 2015. "TRAViz: A visualization for Variant Graphs". *Digital Scholarship in the Humanities*, 30(1), pp. 83–99.

Kluyver, Thomas, et al. 2016. "Jupyter Notebooks — A Publishing Format for Reproducible Computational Workflows". In F. Loizides and B. Schmidt (eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. N.p.: IOS Press, pp. 87–90.

Macé, Caroline, et al. 2015. "Textual criticism and text editing". In Alessandro Bausi, et al. (eds.), *Comparative Oriental Manuscript Studies: An Introduction*. Hamburg: Tredition, pp. 321–466.

Ott, W. 1991. "TUSTEP". In Heinrich Best (ed.), *Computers in the Humanities and the Social Sciences: Achievements of the 1980s, Prospects for the 1990s.Proceedings of the Cologne Computer Conference 1988 Uses of the Computer in the Humanities and Social Sciences held at the University of Cologne, September 1988.* Köln: De Gruyter, pp. 432–37.

Pasquali, Giorgio. 1952. *Storia della Tradizione e Critica del Testo*. Firenze: Felice Le Monnier.

Pithoeus, Petrus, ed. 1594. *M. Fab. Quintiliani Declamationes, quae ex*

CCCLXXXVIII. supersunt, CXLV. ex vetere exemplari restitutae. Calpurnii Flacci excerptae X. Rhetorum minorum LI. nunc primum editae. Dialogus de oratoribus, sive de caussis coruptae eloquientiae. Ex bibliotheca P. Heidelberg: Hieronymus Commelinus, <http://dx.doi.org/10.3931/e-rara-3278>.

Prebor, Gila. 2013. "New Technologies for the Collation of Hebrew Texts". *Zutot: Perspectives on Jewish Culture*, 10, pp. 53–64.

Robinson, Peter M. W. 1994. "Collate: A Program for Interactive Collation of Large Textual Traditions". In Susan Hockey and Nancy Ide (eds), *Research in Humanities Computing 3*. Oxford: Oxford University Press, pp. 32–45.

Roelli, Philip. 2014. "Petrus Alfonsi, or: On the Mutual benefit of Traditional and Computerised stemmatology". In Tara Andrews and Caroline Macé (eds.), *Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches*. Turnhout: Brepols, pp. 43-68.

Roos, Teemu and Yuan Zou. 2011. "Analysis of Textual Variation by Latent Tree Structures". Proceedings — IEEE 11th International Conference on Data Mining. N.p. IEEE Computer Society Digital Library, pp. 567–576, <https://doi.ieeecomputersociety.org/10.1109/ICDM.2011.24>.

Schmidt, Desmond and Robert Colomb. 2009. "A Data Structure for Representing multi-version Texts Online". *International Journal of Human-Computer Studies* 67(6), pp. 497–514.

Spadini, Elena. 2015. "Annotating document changes". In *Proceedings of the 3rd International Workshop on (Document) Changes: Modeling, Detection, Storage and Visualization Lausanne, Switzerland — September 08 - 08, 2015*, pp. 23–26.

TEI Consortium. 2016. 3.4.2 Regularization and Normalization. TEI P5. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/CO.html#COEDREG> [Accessed March 3, 2017].

Trovato, Paolo. 2014. Everything You Always Wanted to Know about Lachmann's Method. A Non-Standard Handbook of Genealogical Textual Criticism in the Age of Post-Structuralism, Cladistics, and Copy-Text. Padova: libreriauniversitaria.it edizioni.

VanderPlas, Jake. 2016. *Python Data Science Handbook*. Beijing: O'Reilly Media.

West, Martin L. 1973. *Textual Criticism and Editorial Technique, applicable to Greek and Latin Texts.* Stuttgart: Teubner.